| Model | Embedding Similarity Search | | | | |
|---|---|---|---|---|---|
| | R@3 | R@5 | R@10 | R@20 | R@50 |
| Starmie | 0.421 | 0.540 | 0.635 | 0.730 | 0.825 |
| Ours | **0.722** | **0.833** | **0.897** | **0.944** | **0.992** |

| Model | Column Type Annotation | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | k=10 | | | k=20 | | | k=50 | | |
| | R@1 | R@3 | R@5 | R@1 | R@3 | R@5 | R@1 | R@3 | R@5 |
| Starmie | 0.429 | – | – | 0.460 | – | – | 0.405 | – | – |
| Ours | **0.802** | **0.881** | **0.897** | 0.619 | 0.833 | 0.889 | 0.722 | 0.762 | 0.881 |

Table 1: Comparison of Recall at different cutoffs (R@k) for Starmie and our data augmentation method on the ARPA schema matching task on 700+ GDC variable names. ”Embedding Similarity Search” refers to using embeddings from a pre-trained language model fine-tuned using contrastive learning for similarity searches to retrieve the top-k entries. ”Column Type Annotation” (CTA) involves utilizing the top-k entries retrieved by the language model for column type annotation, comparing precision at top 1, 3, and 5 matches for label set sizes 10, 20, and 50.